Lee *et al. BMC Medical Genomics* 2020, **13**(Suppl 6):81
https://doi.org/10.1186/s12920-020-00736-7

BMC Medical Genomics

RESEARCH Open Access

Constructing cancer patient-specific and group-specific gene networks with multi-omics data

Wook Lee[1], De-Shuang Huang[2] and Kyungsook Han[1*]

From 15th International Symposium on Bioinformatics Research and Applications (ISBRA '19)
Barcelona, Spain. 3-6 June 2019

Abstract

Background: Cancer is a complex and heterogeneous disease with many possible genetic and environmental causes. The same treatment for patients of the same cancer type often results in different outcomes in terms of efficacy and side effects of the treatment. Thus, the molecular characterization of individual cancer patients is increasingly important to find an effective treatment. Recently a few methods have been developed to construct cancer sample-specific gene networks based on the difference in the mRNA expression levels between the cancer sample and reference samples.

Methods: We constructed a patient-specific network with multi-omics data based on the difference between a reference network and a perturbed reference network by the patient. A network specific to a group of patients was obtained using the average change in correlation coefficients and node degree of patient-specific networks of the group.

Results: In this paper, we present a new method for constructing cancer patient-specific and group-specific gene networks with multi-omics data. The main differences of our method from previous ones are as follows: (1) networks are constructed with multi-omics (mRNA expression, copy number variation, DNA methylation and microRNA expression) data rather than with mRNA expression data alone, (2) background networks are constructed with both normal samples and cancer samples of the specified type to extract cancer-specific gene correlations, and (3) both patient individual-specific networks and patient group-specific networks can be constructed. The results of evaluating our method with several types of cancer show that it constructs more informative and accurate gene networks than previous methods.

Conclusions: The results of evaluating our method with extensive data of seven cancer types show that the difference of gene correlations between the reference samples and a patient sample is a more predictive feature than mRNA expression levels and that gene networks constructed with multi-omics data show a better performance than those with single omics data in predicting cancer for most cancer types. Our approach will be useful for finding genes and gene pairs to tailor treatments to individual characteristics.

Keywords: Individual-specific gene network, Group-specific gene network, Cancer, Multi-omics data

*Correspondence: khan@inha.ac.kr
[1] Department of Computer Engineering, Inha University, 22212 Incheon, South Korea
Full list of author information is available at the end of the article

Background

For the past years, we have witnessed the rapid development of targeted cancer therapy. Targeted therapies for cancer work by targeting specific genes, proteins or tissues that contribute to cancer growth and survival. Many targeted therapies are effective only for patients with specific genetic alterations (known as driver mutations) that help cancer cells form and grow [1, 2]. Thus, identifying genetic mutations specific to individual patients is of utmost importance to determine targeted therapies that can effectively cure cancer patients while minimizing side effects [3].

Motivated by a massive amount of data generated by high-throughput technologies, several cancer studies used gene networks to explore gene expression characteristics [4–8]. However, constructing a patient-specific gene network with a single sample obtained from a patient is difficult because a gene network requires many samples to compute gene-gene relations.

Recently a few methods have been proposed to construct cancer sample-specific gene networks based on the difference in the mRNA expression levels between the cancer sample and reference samples. For example, Liu et al. [9] proposed a method to construct a sample-specific network by computing the difference between a reference network from multiple reference samples and a network perturbed by a new sample. However, a slight change to the reference samples can result in a significantly different sample-specific network for the same sample due to the small number of reference samples. Furthermore, their sample-specific networks cannot reflect post-translational modification and epigenetics because the networks are built using mRNA expression data only.

This paper presents a new method for constructing cancer patient-specific and group-specific gene networks with multi-omics data using a sample-specific network and network propagation method. Network propagation strategies are widely used in recent cancer-related research. Li et al. [10] presented a synergy prediction algorithm using network propagation and predicted the drug synergy in various cancers. Zhang et al. [11] introduced a propagation algorithm, which learns the mutated subnetworks underlying tumor subtypes using a supervised approach and classified tumors to known subtypes on breast and glioblastoma tumors. Peng et al. [12] identified bladder cancer-related genes by propagating information from seed genes to candidate genes. The primary focus of our method is to construct a gene correlation network specific to cancer with multi-omics data. Thus, it is different from a typical gene co-expression network that represents co-expression relations between genes from mRNA expression data. Our gene network is not a gene regulatory network because our network does not show regulatory relations between genes.

The main differences of our method from previous ones are as follows: (1) networks are constructed with multi-omics (mRNA expression, copy number variation, DNA methylation and microRNA expression) data rather than with mRNA expression data alone, (2) background networks are constructed with both normal samples and cancer samples of the specified type to extract cancer-specific gene correlations, and (3) both patient individual-specific networks and patient group-specific networks can be constructed. As shown later in this paper, the results of evaluating our method with several types of cancer show that it constructs more informative and accurate target-specific networks than previous methods.

Methods

At the top level, our method consists of the following steps: (1) data processing, (2) constructing individual-specific gene networks, and (3) constructing a group-specific gene networks. A high-level description of the method is given in Fig. 1.

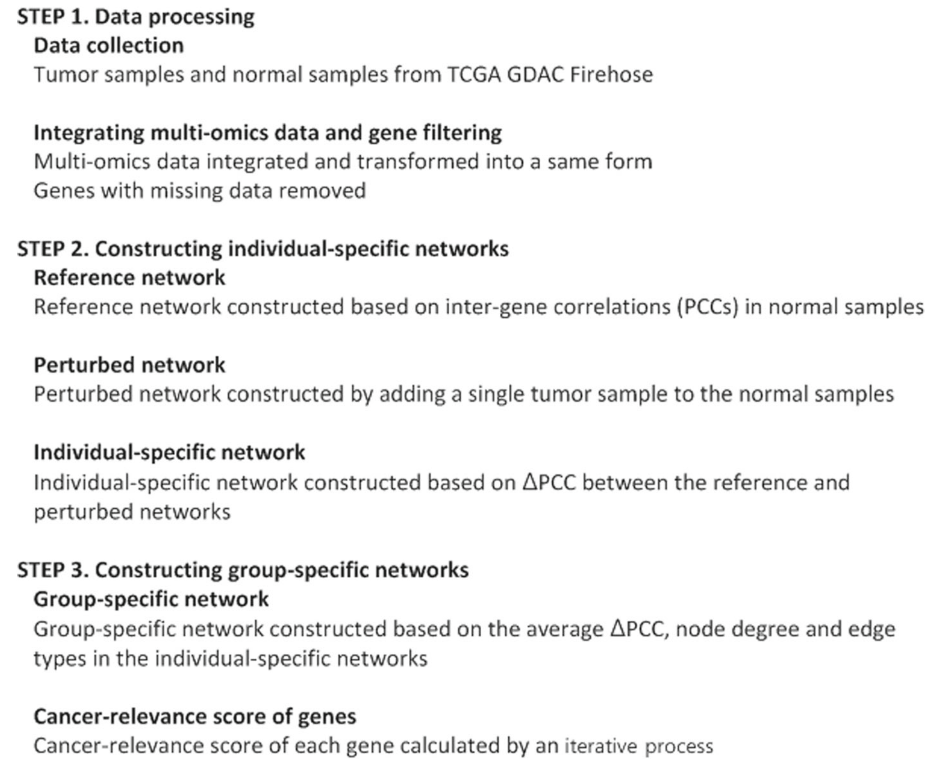
Data collection and preprocessing

From the Broad Institute TCGA GDAC Firehose [13], we obtained multi-omics data of cancer samples of seven types: breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), pan-kidney cohort (KIPAN), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC).

The multi-omics data used in this study include mRNA expression (mRNAseq), copy number variation (CNV), DNA methylation and mature miRNA expression (miRseq) data. The mRNAseq data were processed using quartile normalized RSEM [14] and then log2-transformed. The segmented CNV data were converted to gene-level data using the Ensembl API [15] and the CNTools package [16] of Bioconductor. The methylation data were filtered to select the probe with the mean signal values for each gene. The miRseq data were processed by RPM and log2-transformed. mRNAs and miRNAs that were not expressed in more than 10% of the total samples were excluded in further analysis. Missing expression values of mRNAs and miRNAs were replaced by the smallest positive normalized floating-point number (realmin) of MATLAB. The number of samples and genes used in this study are available in Additional file 1.

Individual-specific gene network

In each group of tumor samples and normal samples, we first computed gene-gene relations by the Pearson correlation coefficient (PCC), selected highly correlated gene pairs (i.e., those with $|PCC| > 0.8$), and constructed two sample networks, one for each group. From the tumor sample network, we removed edges common to both



STEP 1. Data processing

Data collection

Tumor samples and normal samples from TCGA GDAC Firehose

Integrating multi-omics data and gene filtering

Multi-omics data integrated and transformed into a same form

Genes with missing data removed

STEP 2. Constructing individual-specific networks

Reference network

Reference network constructed based on inter-gene correlations (PCCs) in normal samples

Perturbed network

Perturbed network constructed by adding a single tumor sample to the normal samples

Individual-specific network

Individual-specific network constructed based on ΔPCC between the reference and perturbed networks

STEP 3. Constructing group-specific networks

Group-specific network

Group-specific network constructed based on the average ΔPCC, node degree and edge types in the individual-specific networks

Cancer-relevance score of genes

Cancer-relevance score of each gene calculated by an iterative process

Fig. 1 Overview of constructing an individual-specific network and a group-specific network with multi-omics data

tumor and normal sample networks and obtained a template reference network for cancer (Fig. 2a). The template reference network consists of highly-correlated gene pairs that are specific to cancer.

With n reference samples, which may be different from tumor samples used in the template network, we computed PCC for every pair of genes in the template reference network and constructed a reference network for the reference samples. For a patient of interest, we constructed a network, which is a perturbed network by adding a single sample of the patient to the n reference samples. A patient-specific network was obtained by subtracting the reference network from the perturbed network.

$$\Delta PCC = |PCC_{n+1}| - |PCC_n| \tag{1}$$

We computed the difference in the absolute value of PCC between the perturbed reference network and reference network by Eq. 1. We also carried out a Z-test of $PCC_{n+1} - PCC_n$ by Eq. 2. For a large n, we can approximate the mean (μ) and standard deviation (σ) of $PCC_{n+1} - PCC_n$ as 0 and $(1 - PCC_n^2)/(n-1)$, respectively [9].

$$\begin{aligned} PCC_{change} &= PCC_{n+1} - PCC_n \\ Z - score &= \frac{PCC_{change} - \mu(PCC_{change})}{\sigma(PCC_{change})} = \frac{PCC_{change}}{\frac{1-PCC_n^2}{n-1}} \end{aligned} \tag{2}$$

The edges of the patient-specific network were classified into four types [9]: (1) correlation-gained edges for gene pairs whose PCCs are increased from the reference network to the patient-specific network, (2) correlation-lost edges for gene pairs whose PCCs are decreased from the reference network to the patient-specific network, (3) correlation-reversed edges for gene pairs whose signs of PCCs are changed from positive to negative or negative to positive, and (4) correlation-invariant edges for gene pairs with little change in PCCs between the reference and patient-specific networks (i.e., those with |Z-score| $<$ 1) (Fig. 2b).

The edges were classified in the following way. We first selected gene pairs with |Z-score| $<$ 1 as correlation-invariant type, and then selected gene pairs which have different signs of PCCs between the reference network and the patient-specific network as correlation-reversed type. The remaining gene pairs were classified into either correlation-gained or correlation-lost type depending on whether their PCCs are increased (correlation-gained) or decreased (correlation-lost) from the reference network to the patient-specific network. Thus, |Z-score| $\geq$ 1 in both correlation-gained and correlation-list gene pairs.

Group-specific gene network

A group-specific gene network is useful when analyzing a large number of patient-specific gene networks. After constructing patient-specific gene networks, we obtained

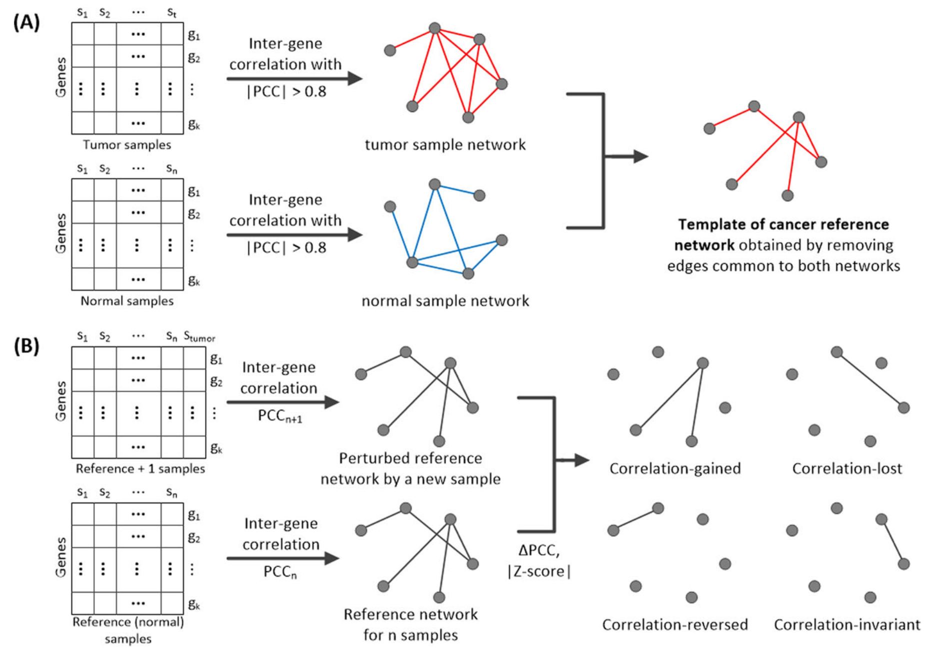

(A)
s1 s2 ··· st
g1
g2
gk
Genes
Tumor samples
Inter-gene correlation with |PCC| > 0.8
tumor sample network
s1 s2 ··· sn
Genes
Normal samples
Inter-gene correlation with |PCC| > 0.8
normal sample network
Template of cancer reference network obtained by removing edges common to both networks
(B)
s1 s2 ··· sn stumor
Genes
Reference + 1 samples
Inter-gene correlation PCCn+1
Perturbed reference network by a new sample
s1 s2 ··· sn
Genes
Reference (normal) samples
Inter-gene correlation PCCn
Reference network for n samples
ΔPCC, |Z-score|
Correlation-gained
Correlation-lost
Correlation-reversed
Correlation-invariant

Fig. 2 Process of constructing a patient-specific gene network. **a** template of the cancer reference network obtained by removing edges common to both networks. **b** patient-specific gene network and four types of edges in the network. Edges with a |, Z-score| , < 1 represent correlation-invariant gene pairs, and edges with |, Z-score| , ≥ 1 and different signs of PCC_n and PCC_{n+1} represent correlation-reversed gene pairs. Edges with |Z-score| ≥ 1 and $\Delta PCC > 0$ are correlated-gained gene pairs, and those with |Z-score| ≥ 1 and $\Delta PCC < 0$ are correlation-lost gene pairs

a gene network specific to a group of patients based on the average ΔPCC and node degree of the patient-specific networks (Fig. 3). If the dominant type for a particular edge is 'correlation-gained' (positive ΔPCC) in the patient-specific networks, the edge is represented in red in the group-specific network. In contrast, if the dominant type for a particular edge is 'correlation-lost' (negative ΔPCC) in the patient-specific networks, the edge is represented in blue in the group-specific network. In the group-specific network, only the dominant type is shown for each edge. If non-dominant types are shown in addition to the dominant type for each edge, the network becomes cluttered and unreadable. The node size of a group-specific gene network is proportional to the average degree of the node.

Integration of multi-omics data

To integrate multi-omics data, we first computed inter-gene correlations by PCC with four different types of single omics data (mRNA expression, CNV, DNA methylation and miRNA expression) separately, and selected significant inter-gene correlations only. In mRNA expression, CNV and DNA methylation data, we select the top 1% $|PCC|$ with p-value <0.01. In miRNA expression data, we selected the top 5% $|PCC|$ with p-value <0.01 due to a smaller number of miRNAs in the data. The inter-gene correlations selected in each single omics data are represented in four correlation matrices (M_{expr}, M_{CNV}, M_{methyl} and M_{miRNA}) and normalized.

Using the protein-protein interactions (PPIs) of the STRING database [17], we constructed separate weighted networks from each omics data by Eq. 3. In Eq. 3, W_{expr}, W_{CNV} and W_{methyl} denote the weighted networks, and PPI_{expr}, PPI_{CNV} and PPI_{methyl} are subnetworks of a PPI network consisting of genes present in each omics data. Since the PPI network does not contain information on miRNA, a weighted network for miRNA was not constructed.

$$\begin{aligned} W_{expr} &= 1 - \left(1 - M_{expr}\right) \times \left(1 - PPI_{expr}\right) \\ W_{CNV} &= 1 - \left(1 - M_{CNV}\right) \times \left(1 - PPI_{CNV}\right) \\ W_{methyl} &= 1 - \left(1 - M_{methyl}\right) \times \left(1 - PPI_{methyl}\right) \end{aligned} \tag{3}$$

We then integrated the multi-omics data by linear regression using Eq. 4 [12]. In Eq. 4, Y_i, X_i^{CNV}, X_i^{methyl} and X_{ij}^{miRNA} denote gene i's expression level, CNV level, methylation level, and miRNA regulator expression level, respectively. β_i^{CNV} and β_i^{methyl} denote the regression coefficients of gene i's expression level on CNV and methylation, respectively. β_{ij}^{miRNA} is the regression coefficient of gene i's expression level on its miRNA regulator j's expression level.

$$Y_i = \beta_i^{CNV} X_i^{CNV} + \beta_i^{methyl} X_i^{methyl} + \sum_{j=1}^{n} \beta_{ij}^{miRNA} X_{ij}^{miRNA} + \epsilon \tag{4}$$

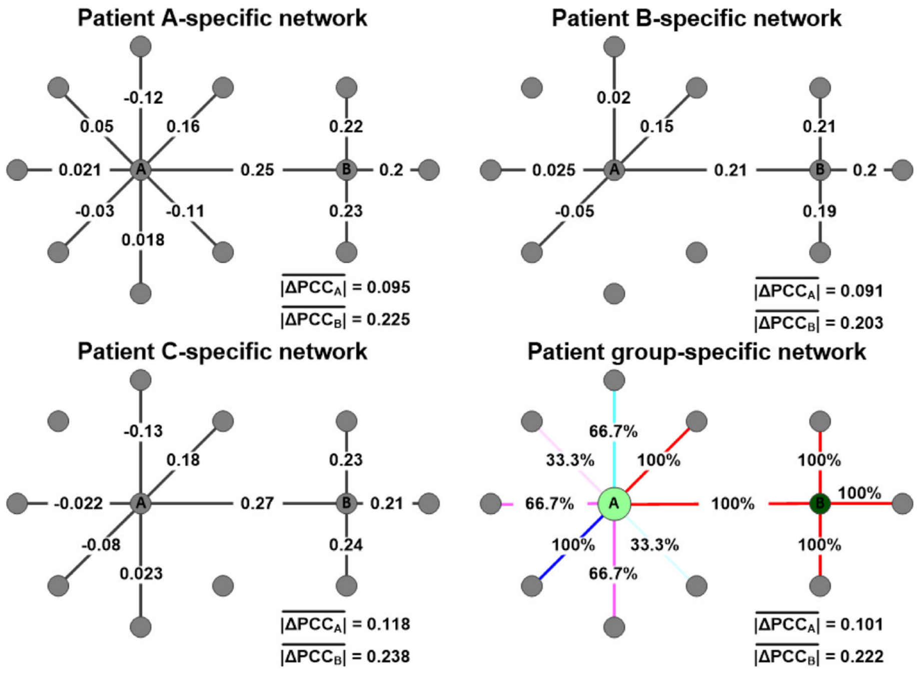


Fig. 3 Process of constructing a group-specific gene network from multiple patient-specific gene networks. In the group-specific gene network, the node size and node color are proportional to the average degree and average $|\Delta PCC|$ of the node, respectively. If the dominant type for a particular edge is 'correlation-gained' (positive ΔPCC) in the patient-specific networks, the edge is represented in red in the group-specific network. If the dominant type for a particular edge is 'correlation-lost' (negative ΔPCC) in the patient-specific networks, the edge is represented in blue in the group-specific network

From the regression coefficients and the weighted networks, a weight matrix W was derived and normalized into $\overline{W}$ (Eqs. 5 and 6). The weight matrix W is symmetric, so $W_{ij} = W_{ji}$. W_{11}, W_{22}, W_{33} and W_{44}, represent W_{expr}, W_{CNV}, W_{methyl} and M_{miRNA}, respectively. The submatrices W_{21} and W_{31} contain regression coefficients β_i^{CNV} and β_i^{methyl} for every gene i, respectively. W_{41} represents β_{ij}^{miRNA}. The submatrices W_{32}, W_{42} and W_{43} are empty.

$$W = \begin{bmatrix} W_{11} & W_{12} & W_{13} & W_{14} \\ W_{21} & W_{22} & W_{23} & W_{24} \\ W_{31} & W_{32} & W_{33} & W_{34} \\ W_{41} & W_{42} & W_{43} & W_{44} \end{bmatrix} \tag{5}$$

$$\overline{W}(i,j) = W(i,j) / \sqrt{\sum_{k=1}^{m} W(i,k) \times \sum_{k=1}^{m} W(k,j)} \tag{6}$$

In network propagation, seed genes have greater impact than non-seed genes on their neighbors. Thus, only the genes with a high average ΔPCC were selected as seed genes for a group-specific network, and their miRNAs regulators extracted from miRTarBase [18] were used as seed miRNAs. We calculated the cancer-relevance S^t of each gene to reflect the effect of the seed genes and miRNAs on neighbors. The initial score D was calculated by Eq. 7 and updated iteratively by Eq. 8 [12]. In this iterative process, the influence of the seed is propagated to the neighbors until a mean squared error of S^t and S^{t-1} $\leq 1 \times 10^{-5}$.

$$D_v = \begin{cases} \dfrac{n_v}{N_v} & \text{if } v \text{ is a non-seed \& } N_v \geq \alpha \\ e^{N_v - \alpha} \times \dfrac{n_v}{N_v} & \text{if } v \text{ is a non-seed \& } N_v < \alpha \\ 1 & \text{if } v \text{ is a seed} \end{cases} \tag{7}$$

$$S^t = \lambda \times S^{t-1} \times \overline{W} + (1-\lambda) \times D, \text{where} S^1 = D \tag{8}$$

where N_v is the number of neighbors of node v, and n_v is the number of seeds in the neighbors. The parameter α, which is a threshold for N_v, was set to 50 and λ was set to 0.2 [12]. Genes with the top 10% S^t were used in finding cancer-related genes and in classifying tumor samples and normal samples.

Results

Patient-specific and group-specific gene networks

In this study, we constructed 2,400 patient-specific gene networks for seven cancer types (Additional file 1). For each cancer type, we also constructed group-specific gene networks. As an example, Fig. 4 shows a group-specific gene network derived from 300 lung squamous cell carcinoma (LUSC) patients.

There are three distinct subnetworks in the network for the LUSC group. The subnetwork enclosed in box A of Fig. 4 contains many hub genes (large green nodes). The subnetwork in box B consists of correlation-gained edges (dark red edges), whereas the subnetwork in box C contains many correlation-lost edges (dark blue edges).

Comparison of multi-omics data and single-omics data

We performed leave-one-out cross validation (LOOCV) to evaluate cancer-relevance score S^t of a gene and the contribution of multi-omics data to finding cancer-related genes. For comparison, the cancer-relevance scores were computed with multi-omics data and single omics data separately. Each seed gene was regarded as a non-seed and a new cancer-relevance score was calculated for the gene. Seed genes and non-seed genes were considered as positive and negative, respectively. Seed genes included in the top n genes were considered as true positives, and seed genes not included in the top n genes were considered as false negatives. Similarly, non-seed genes included in

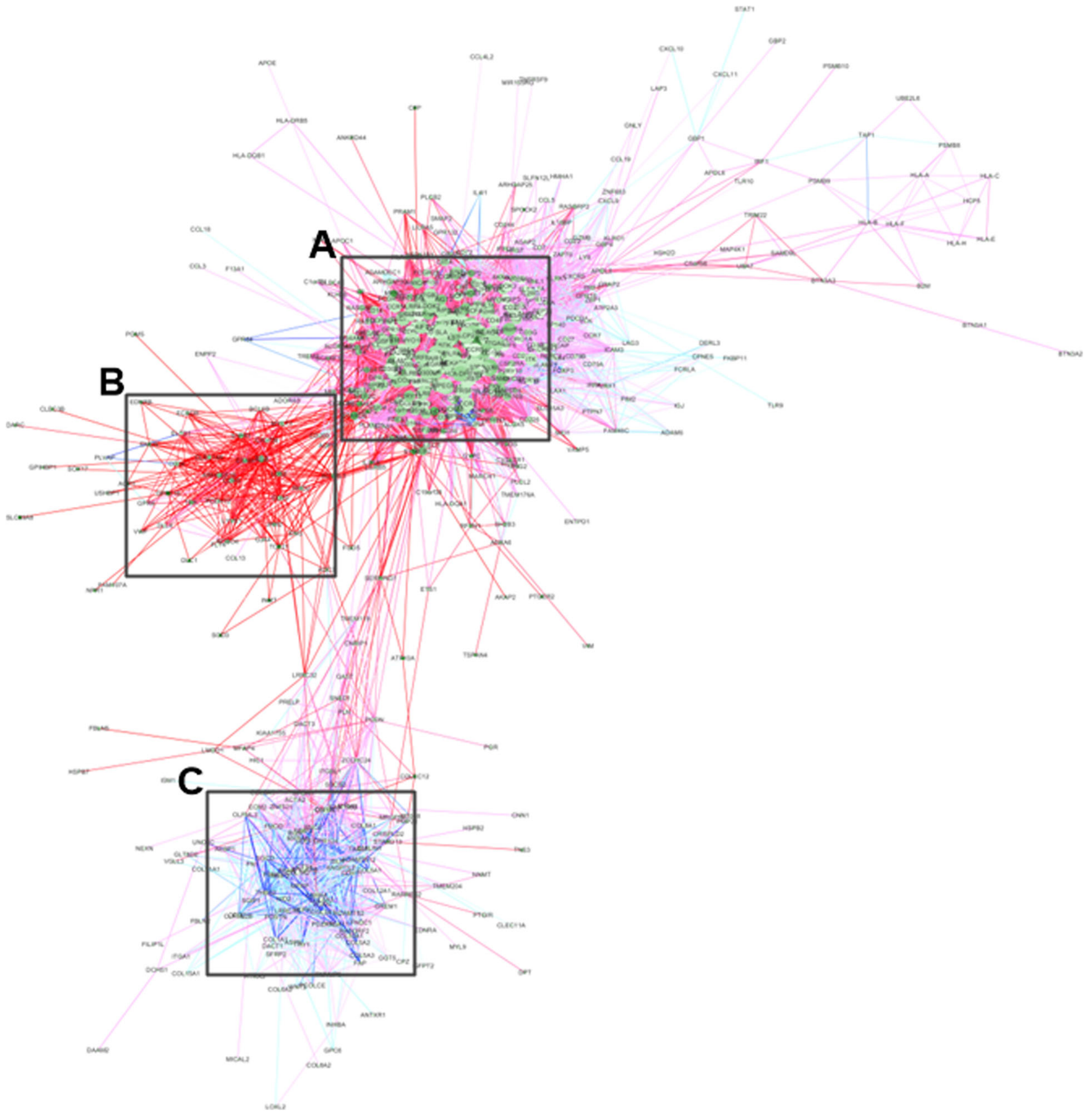


Fig. 4 Group-specific gene network for 300 lung squamous cell carcinoma (LUSC) patients. **a** subnetwork of multiple hub genes (large green nodes). **b** subnetwork of correlation-gained edges (dark red edges). **c** subnetwork with many correlation-lost edges (dark blue edges)

the top *n* genes and non-seed genes not included in the top *n* genes were considered as false positives and true negatives, respectively.

We carried out LOOCV with different ratios of seed genes to non-seed genes. Figure 5 shows the receiver operating characteristic (ROC) curve and the area under the curve (AUC) of LOOCV of the cancer relevance of genes on data of 400 breast cancer samples with various seed ratios ranging from 0.01 to 0.09 (Enlarged plots of Fig. 5 are available in Additional file 2). For comparative purposes, we also computed the cancer relevance of genes with single omics data. As shown in Fig. 5, multi-omics data consistently exhibited better performance than single omics data with any seed ratio between 0.01 to 0.09. For later analysis, the seed ratio was set to 0.05 by default. The average ΔPCC and class label of each gene are available in Additional file 3.

Indeed, the superiority of multi-omics data over single omics data in determining the cancer relevance score of genes was observed in all seven types of cancer (Additional file 4). In seven types of cancer, the cancer relevance score of genes computed with multi-omics data exhibited a good performance (AUC = 0.896 ∼0.942). The cancer relevance score of genes computed with mRNA expression data showed the second best performance (AUC = 0.761 ∼0.878). In particular, the cancer relevance score computed with mRNA expression data showed a very similar performance to that with multi-omics in breast cancer (BRCA). The performance of the cancer relevance score computed with CNV (AUC = 0.591 ∼0.786) and DNA methylation data (AUC = 0.581 ∼0.817) alone was lower than that with mRNA expression data (AUC = 0.761 ∼0.878).

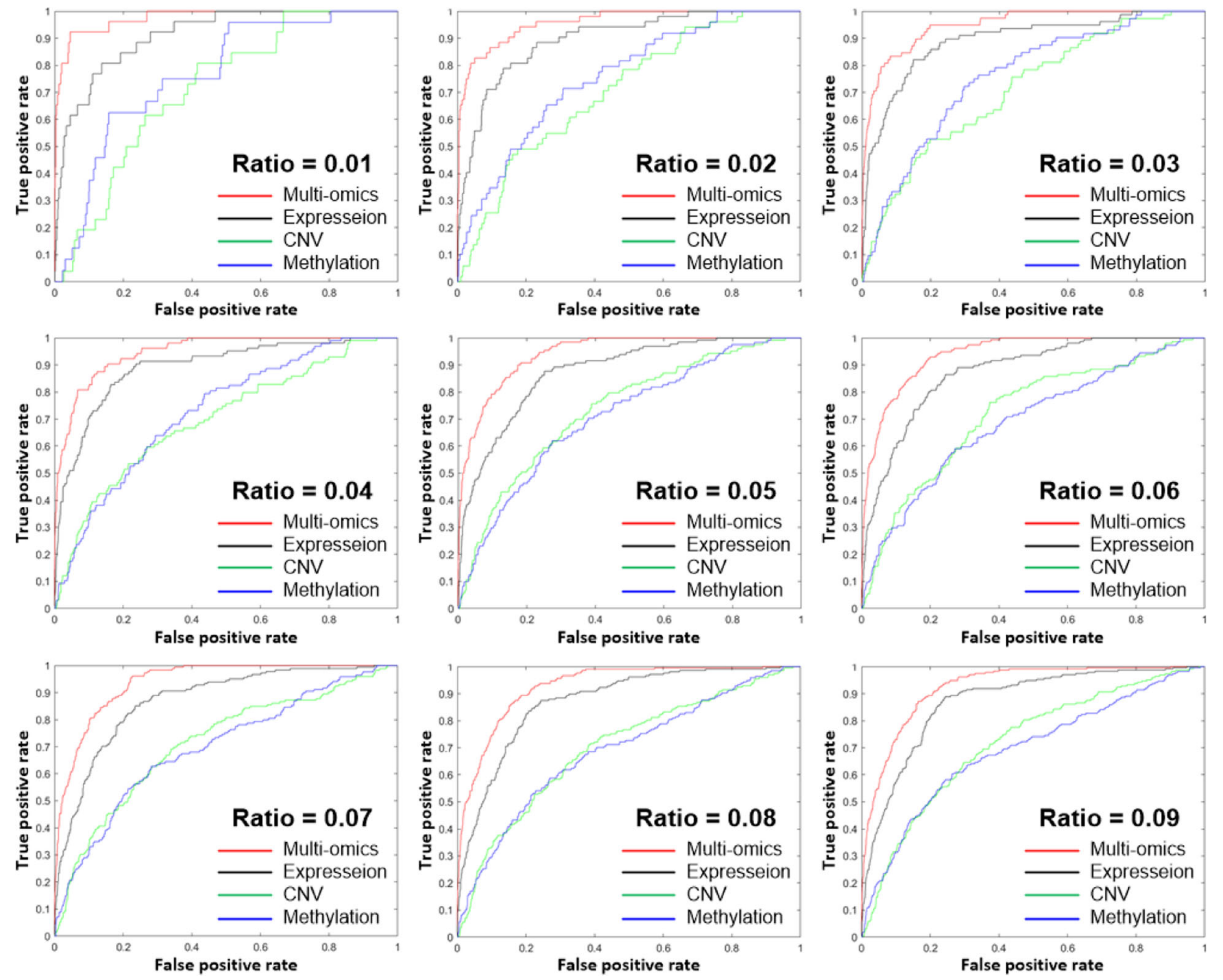


Fig. 5 ROC curves of the leave-one-out cross validation of the cancer relevance score S^t of genes with different ratios of seed genes. 400 breast cancer samples were used in the leave-one-out cross validation. The performance of multi-omics data is always better than that of single omics data

Evaluation of gene correlations and background networks

Many network-based approaches to cancer research have focused on finding genes that show differential expressions between tumor samples and normal samples. Gene-gene correlations (i.e., inter-gene correlations) may be more helpful than individual genes because inter-gene correlations depend on the expression of neighbor genes in a gene regulatory network. To compare the effect of using individual genes to that of inter-gene correlations (i.e., ΔPCC), we constructed a support vector machine (SVM) model for classifying cancer samples and normal samples. The SVM model was implemented using C-SVC and RBF kernel, and the parameter values of the model were determined by the grid search algorithm. mRNA expression levels and ΔPCCs were used as features of the SVM models. For rigorous validation, the test data used in testing the models were not used in training them (Additional file 1).

As shown in Fig. 6a, ΔPCC showed a better performance than mRNA expression levels for six cancer types except LUSC. The classification model with ΔPCC showed MCC above 0.9 in six cancer types except HNSC.

We also examined the effect of different background networks on individual-specific networks. In the work by Liu et al. [9], PPI data with high confidence scores in the STRING database were used to construct a background network. However, the PPI data of STRING does not reflect cancer type-specific characteristics. Figure 6b shows the performance of the classification model with two different background networks: background network from PPI data of STRING (the approach by Liu et al.) and cancer background network (our approach). ΔPCC was used as a feature of the classification model. Except for COAD, the performance of the classification model with the cancer background network was better than the model with the STRING reference background network. In particular, the classification model showed a significant difference for breast cancer (BRCA) (MCC of 0.992 vs. MCC of 0.841). Detailed results of the classification model are available in Additional file 5.

Discussion

In the analysis of finding cancer-related genes and gene pairs, we focused on a subnetwork of genes with a ΔPCC. Table 1 shows the top 10 genes with a high average ΔPCC in each group-specific network of seven cancer types. In breast invasive carcinoma (BRCA), FAM171A1 showed the highest average ΔPCC in the group-specific network. FAM171A1 is known as a potential biomarker in triple-negative breast cancer [19]. FOXC1 is involved in tumor development and metastasis and associated with poor prognosis in basal-like breast cancer [20]. IL-33 is overexpressed in various cancers and the serum concentration of IL-33 is a valuable indicator of poor prognosis in breast cancer. [21]. MAMDC2 is significantly correlated with disease-free survival of breast cancer patients [22]. MTERFD1 is closely related to breast cancer recurrence

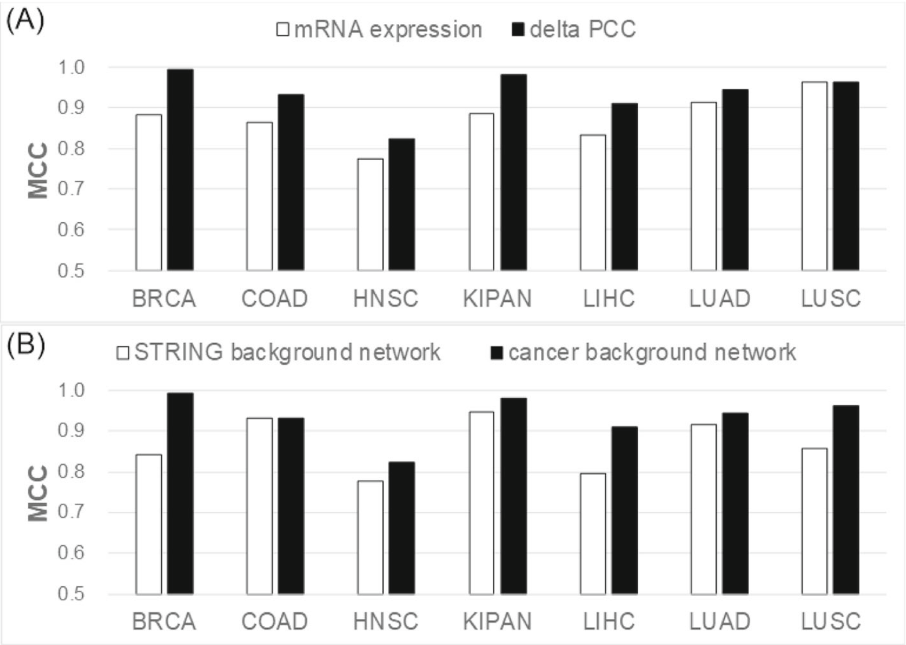


Fig. 6 Results of evaluating features and background networks by a validation set. **a** Comparison of mRNA expressions of genes and ΔPCC of gene pairs. **b** Comparison of the cancer background network with the background network from PPI data

Table 1 Top 10 genes with a high average ΔPCC in a group-specific network for seven cancer types

BRCA	**COAD**	**HNSC**	**KIPAN**	**LIHC**	**LUAD**	**LUSC**
FAM171A1	GFRA2	CYP2J2	TFCP2L1	SLC19A3	CLDN18	NSUN2
FOXC1	SCNN1B	BARX2	KCNQ1	ECM1	ADAMTS8	CCT5
IL33	DDX27	ZNF135	ARL15	CYP2B6	PECAM1	FBXO45
MAMDC2	RNPS1	PPFIA1	KCTD1	FBP1	SFTPA1	GPR116
MTERFD1	UBE2I	PARL	OAZ2	GPAA1	GIMAP6	SLC39A8
HOXA7	MDFIC	FADD	TMEM45B	F9	AKR1C1	FXR1
CTTNBP2	CTNNBL1	CST6	HPCAL1	PAH	MME	INMT
ZNF204P	CDK5RAP1	COBL	TMEM91	AGXT2	ATAD2	VEPH1
JAM3	ESF1	ORAOV1	SEMA5B	HSD17B6	CYP3A5	CRTAM
FREM1	TRMT6	XPO7	EGLN3	RNASE4	CHAF1B	WDR53

[23] and HOXA7 plays a critical role in regulating the proliferation of ER-positive cancer cells [24].

In colon adenocarcinoma (COAD), GFRA2 showed the highest average ΔPCC in the group-specific network. It is known to be crucial for enteric neuron survival [25]. SCNN1B and DDX27 are significantly related to colorectal cancer [26, 27]. No direct relation of RNPS1 with colorectal cancer is known, but RNPS1 is essential to nonsense-mediated mRNA decay [28] that plays complex functions in cancer [29]. Knockdown of SUMO conjugating enzyme UBE2I (also known UBC9 or E2) inhibits maintenance and self-renewal of colorectal cancer stem cell, while overexpression of UBE2I increases colorectal cancer cell stemness [30].

Among the top 10 genes with a high average ΔPCC in lung adenocarcinoma (LUAD), several genes such as CLDN18, ADAMTS8, PECAM1 and SFTPA1 have been known to be associated with LUAD in previous studies [31–33]. No direct relation of NSUN2 and SLC39A8 with lung squamous cell carcinoma (LUSC) has been known so far. However, recent studies [34, 35] reported that NSUN2 is correlated with survival in other types of squamous cell carcinomas. Gao et al. also showed that the epigenetic silencing of SLC39A8 expression by DNA methylation is involved in the acquisition of resistance against cadmium in lung cells [36] and the relation between cadmium and lung cancer has received much attention [37]. Many other genes in Table 1 found in the group-specific networks for head and neck squamous cell carcinoma (HNSC), pan-kidney cohort (KIPAN) and liver hepatocellular carcinoma (LIHC) are also directly or indirectly related to cancer.

In addition to individual genes, we identified gene pairs of the same type (i.e., either correlation-gained or correlation-lost in most patient-specific networks of the same type). Table 2 shows the most frequent gene pairs in 400 breast cancer samples. The most frequent gene pairs in other types of cancer are listed in Additional file 6. It is interesting to note that all the gene pairs shown in Table 2 include at least one gene in the gene pair MAMDC2-HOXA7 and that they are correlation-gained edges in the group-specific network for breast cancer. Figure 7 shows a subnetwork containing MAMDC2 and HOXA7 in the group-specific network of breast cancer. The subnetwork was obtained by selecting the edges for which the proportion of the same edge type (i.e., correlation-gained or lost) is above 90% in the total individual-specific networks of breast cancer patients. It is interesting to note that all the gene pairs in Table 2 are included in the subnetwork.

To date, the actual role of the MAMDC2 gene in cancer is not clear, but Meng et al. [22] reported MAMDC2 as one of three genes (MAMDC2, TSHZ2, and CLDN11) that are significantly correlated with disease-free survival of breast cancer patients. MAMDC2 is known as a target of miR-196a in head and neck squamous cell carcinoma [38]. As a member of the family of homeobox genes, HOXA7 is associated with cell proliferation, nerve invasion, distant metastasis and degree of tumor differentiation in several cancers [24, 39–42]. HOXA7 is regulated

Table 2 The most frequent gene pairs in 400 breast cancer samples. All the gene pairs are of a correlation-gained type. The genes of Table 1 are shown in bold. The proportion represents the ratio of the gene pairs of the same type (i.e., correlation-gained or lost) to the total number of patient-specific networks

Gene pair	#gene pairs	Proportion of the gene pairs in total cancer samples
MAMDC2-HOXA7	380	95.0%
MAMDC2-CCL14	379	94.8%
MAMDC2-ZNF204P	377	94.3%
MAMDC2-KL	376	94.0%
MAMDC2-SVEP1	376	94.0%
MAMDC2-CORO2B	375	93.8%
HOXA7-MEOX2	372	93.0%
HOXA7-HOXA9	366	91.5%
MAMDC2-SOBP	366	91.5%
MAMDC2-HOXA9	365	91.3%

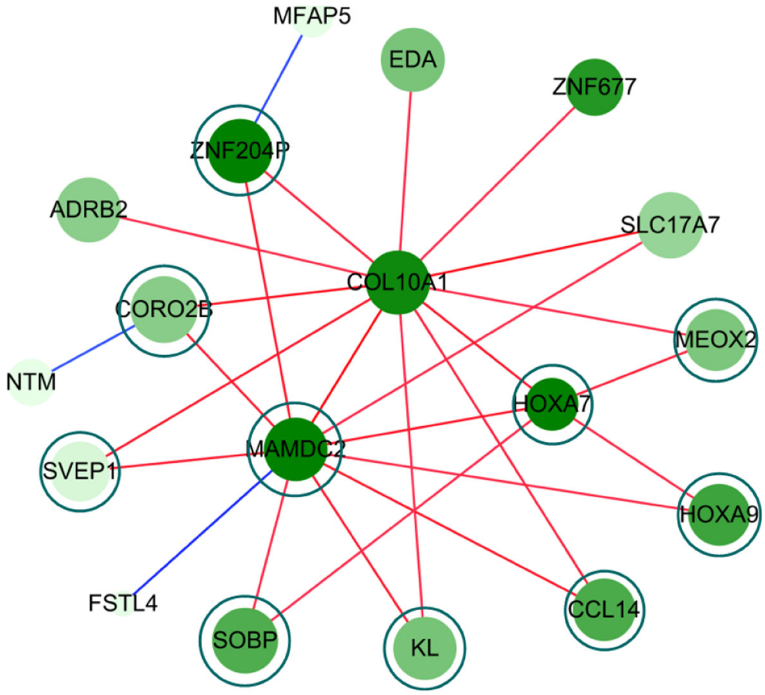


Fig. 7 A subnetwork of the group-specific network of BRCA, which contains MAMDC2 and HOXA7. Genes in the most frequent gene pairs shown in Table 2 are enclosed by a circle

by several miRNAs, including miR-196 [43–45]. Thus, both MAMDC2 and HOXA7 are related with miR-196, but a clear relation among them is to be uncovered.

Conclusion

So far, most approaches to constructing individual-specific gene networks have been constructed based on the differential expressions between a small number of reference samples and a sample of interest. However, such networks cannot reflect post-translational modification and epigenetics and are not reliable because a slight change to the reference samples can result in a significantly different sample-specific network for the same sample.

In this paper, we presented a new approach to constructing cancer patient-specific and group-specific networks with multi-omics data. The main differences of our method from previous ones are as follows: (1) gene networks are constructed with multi-omics (mRNA expression, copy number variation, DNA methylation and microRNA expression) data rather than with mRNA expression data alone, (2) background networks can be constructed with cancer samples of the specified type, and (3) both patient individual-specific networks and patient group-specific networks can be constructed. The results of testing our method with several cancer types showed that it constructs more informative and accurate gene networks than existing methods.

Evaluation of our method with extensive data of seven cancer types showed that changes in gene correlations (ΔPCC) between the reference samples and a patient sample is a more predictive feature than mRNA expression levels and that gene networks constructed with multi-omics data are more powerful than those with single omics data in predicting cancer for most cancer types. More work is required to validate the genes and gene pairs identified in the cancer patient-specific and group-specific networks. However, the method for constructing networks specific to individual patients or patient groups with multi-omics data should be useful aids in determining effective treatments to individual characteristics.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12920-020-00736-7.

Additional file 1: Number of samples and genes in 7 types of cancer.

Additional file 2: ROC curve and AUC of the cancer-relevance score of BRCA by various seed ratios.

Additional file 3: Average ΔPCC and class label of each gene in 7 types of cancer.

Additional file 4: ROC curve of the cancer-relevance score of each cancer type with the seed ratio of 0.05.

Additional file 5: Performance of classification of tumor samples and normal samples.

Additional file 6: Top 10 gene pairs for each cancer type.

Abbreviations

AUC: Area under the curve; BRCA: Breast invasive carcinoma; CNV: Copy number variation; COAD: Colon adenocarcinoma; GDAC: Genome data analysis center; HNSC: Head and neck squamous cell carcinoma; KIPAN: Pan-kidney cohort; LIHC: Liver hepatocellular carcinoma; LOOCV: Leave-one-out cross validation; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; MCC: Matthews correlation coefficient; PCC: Pearson correlation coefficient; PPI: Protein-protein interactions; ROC: Receiver operating characteristic; SVM: Support vector machine; TCGA: The cancer genome atlas

Acknowledgements

The authors thank the anonymous reviewers for the constructive comments that contributed to improving the final version of the paper.

About this supplement

This article has been published as part of BMC Medical Genomics Volume 13 Supplement 6, 2020: Selected articles from the 15th International Symposium on Bioinformatics Research and Applications (ISBRA-19): medical genomics. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-13-supplement-6.

Authors' contributions

WL designed and performed all about this study and prepared the initial manuscript. DH helped integrating multi-omics data of breast cancer. KH supervised the work and wrote the manuscript. All authors read and approved the final manuscript.

Authors' information

Wook Lee is currently working toward his PhD degree at the Department of Computer Engineering, Inha University, Korea. De-Shuang Huang is a director of the Institute of Machines Learning and Systems Biology, Tongji University, China. Kyungsook Han is a professor at the Department of Computer Engineering, Inha University, Korea.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (NRF-2017R1E1A1A03069921) and INHA UNIVERSITY Research Grant. Publication of this article was funded by the NRF grant (NRF-2017R1E1A1A03069921). The funders had no role in the design of the study, data collection and analysis, or writing the manuscript.

Availability of data and materials

Additional files are available at http://bclab.inha.ac.kr/CancerNetwork.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Computer Engineering, Inha University, 22212 Incheon, South Korea. [2]Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, 201804 Shanghai, China.

Received: 28 May 2020 Accepted: 5 June 2020 Published: 27 August 2020

References

1. Widakowich C, de Castro G, de Azambuja E, Dinh P, Awada A. Review: side effects of approved molecular targeted therapies in solid cancers. Oncologist. 2007;12(12):1443–55.
2. Liu S, Kurzrock R. Toxicity of targeted therapy: Implications for response and impact of genetic polymorphisms. Cancer Treat Rev. 2014;40(7): 883–91.
3. Verma M. Personalized medicine and cancer. J Personalized Med. 2012;2(1):1–14.
4. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56–68.
5. Madhamshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. Genome Med. 2012;4(5): 41.
6. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nat Commun. 2014;5:3231.
7. Emmert-Streib F, de Matos Simoes R, Mullan P, Haibe-Kains B, Dehmer M. The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. Front Genet. 2014;5:15.
8. Gov E, Arga KY. Differential co-expression analysis reveals a novel prognostic gene module in ovarian cancer. Sci Rep. 2017;7:4996.
9. Liu X, Wang Y, Ji H, Aihara K, Chen L. Personalized characterization of diseases using sample-specific networks. Nucleic Acids Res. 2016;44(22): e164.
10. Li H, Li T, Quang D, Guan Y. Network propagation predicts drug synergy in cancers. Cancer Res. 2018;78(18):5446–57. https://doi.org/10.1158/0008-5472.can-18-0740.
11. Zhang W, Ma J, Ideker T. Classifying tumors by supervised network propagation. Bioinformatics. 2018;34(13):i484–93.
12. Peng C, Li A, Wang M. Discovery of bladder Cancer-related genes using integrative heterogeneous network modeling of multi-omics data. Sci Rep. 2017;7(1):15639.
13. BITGDA Center. Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run. Broad Institute of MIT and Harvard. 2016. https://doi.org/10.7908/c11g0km9.
14. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinforma. 2011;12:323.
15. Daniel RZ, Premanand A, Wasiu A, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. Ensembl 2018. Nucleic Acids Res. 2018;46(D1):D754–61.
16. Zhang J. CNTools: Convert segment data into a region by sample matrix to allow for other high level computational analyses. R package version 1.38.0. 2018.
17. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, Von Mering C, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41(D1):D808–15.
18. Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, Huang WC, Sun TH, Tu SJ, Lee WH, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. 2018;46(D1):D296–302.
19. Guo J, Gong G, Zhang B. Screening and identification of potential biomarkers in triple-negative breast cancer by integrated analysis. Oncol Rep. 2017;38(4):2219–28.
20. Han B, Bhowmick N, Qu Y, Chung S, Giuliano AE, Cui X. FOXC1: an emerging marker and therapeutic target for cancer. Oncogene. 2017;36: 3957–63.
21. Yang ZP, Ling DY, Xie YH, Wu WX, Li JR, Jiang J, Zheng JL, Fan YH, Zhang Y. The association of serum IL-33 and sST2 with breast cancer. Dis Markers. 2015;2015:. https://doi.org/10.1155/2015/516895.
22. Meng L, Xu Y, Xu C, Zhang W. Biomarker discovery to improve prediction of breast cancer survival: using gene expression profiling, meta-analysis, and tissue validation. OncoTargets Ther. 2016;9:6177–85.
23. Chou H, Yao C, Su S, Lee CY, Hu KY, Terng HJ, Shih YW, Chang YT, Lu YF, Chang CW, et al. Gene expression profiling of breast cancer survivability by pooled cDNA microarray analysis using logistic regression, artificial neural networks and decision trees. BMC Bioinforma. 2013;14:100.
24. Zhang Y, Cheng JC, Huang HF, Leung PC. Homeobox A7 stimulates breast cancer cell proliferation by up-regulating estrogen receptor-alpha. Biochem Biophys Res Commun. 2013;440(4):652–7.
25. Barrenschee M, Böttner M, Hellwig I, Harde J, Egberts JH, Becker T, Wedel T. Site-specific gene expression and localization of growth factor ligand receptors RET, GFRα1 and GFRα2 in human adult colon. Cell Tissue Res. 2013;354(2):371–80.
26. Yun Q, Wong CC, Yu J. IDDF2019-ABS-0203 SCNN1B Functions as a tumor suppressor in colorectal cancer by inhibiting RAS-RAF-MEK-ERK pathway. Gut. 2019;68(Suppl 1):A20.

27. Tang J, Chen H, Wong CC, Liu D, Li T, Wang X, Ji J, Sung JJ, Fang JY, Yu J. DEAD-box helicase 27 promotes colorectal cancer growth and metastasis and predicts poor survival in CRC patients. Oncogene. 2018;37(22):3006–21.
28. Mabin JW, Woodward LA, Patton RD, Yi Z, Jia M, Wysocki VH, Bundschuh R, Singh G. The exon junction complex undergoes a compositional switch that alters mRNP structure and nonsense-mediated mRNA decay activity. Cell Rep. 2018;25(9):2431–46.
29. Popp MW, Maquat LE. Nonsense-mediated mRNA decay and cancer. Curr Opin Genet Dev. 2018;48:44–50.
30. Du L, Li YJ, Fakih M, Wiatrek RL, Duldulao M, Chen Z, Chu P, Garcia-Aguilar J, Chen Y. Role of SUMO activating enzyme in cancer stem cell maintenance and self-renewal. Nat Commun. 2016;7:1–11.
31. Luo J, Chimge NO, Zhou B, Flodby P, Castaldi A, Firth AL, Liu Y, Wang H, Yang C, Marconett CN, et al. CLDN18.1 attenuates malignancy and related signaling pathways of lung adenocarcinoma in vivo and in vitro. Int J Cancer. 2018;143(12):3169–80.
32. Choi GC, Li J, Wang Y, Li L, Zhong L, Ma B, Su X, Ying J, Xiang T, Rha SY, et al. The metalloprotease ADAMTS8 displays antitumor properties through antagonizing EGFR–MEK–ERK signaling and is silenced in carcinomas by CpG methylation. Mol Cancer Res. 2014;12(2):228–38.
33. Abraham V, Cao G, Parambath A, Lawal F, Handumrongkul C, Debs R, Delisser HM. Involvement of TIMP-1 in PECAM-1-mediated tumor dissemination. Int J Oncol. 2018;53(2):488–502.
34. Lu L, Zhu G, Zeng H, Xu Q, Holzmann K. High tRNA transferase NSUN2 gene expression is associated with poor prognosis in head and neck squamous carcinoma. Cancer Investig. 2018;36(4):246–53.
35. Li Y, Li J, Luo M, Zhou C, Shi X, Yang W, Lu Z, Chen Z, Sun N, He J. Novel long noncoding RNA NMR promotes tumor progression via NSUN2 and BPTF in esophageal squamous cell carcinoma. Cancer Lett. 2018;430: 57–66.
36. Gao Y, Xu Y, Wu D, Yu F, Yang L, Yao Y, Liang Z, Lau ATY. Progressive silencing of the zinc transporter Zip8 (Slc39a8) in chronic cadmium-exposed lung epithelial cells. Acta Biochim Biophys Sin. 2017;49(5):444–9.
37. Verougstraete V, Lison D, Hotz P. Cadmium, lung and prostate cancer: a systematic review of recent epidemiological data. J Toxicol Environ Health B Crit Rev. 2003;6(3):227–55.
38. Darda L, Hakami F, Morgan R, Murdoch C, Lambert DW, Hunter KD. The role of HOXB9 and miR-196a in head and neck squamous cell carcinoma. PLoS ONE. 2015;10(4):e0122285. https://doi.org/10.1371/journal.pone.0122285.
39. Duan X, Chen H, Ma H, Song Y. The expression and significance of the HOXA7 gene in oral squamous cell carcinoma. J Oral Sci. 2017;59(3): 329–35.
40. Tang B, Qi G, Sun X, Tang F, Yuan S, Wang Z, Liang X, Li B, Yu S, Liu J, et al. HOXA7 plays a critical role in metastasis of liver cancer associated with activation of Snail. Mol Cancer. 2016;15(1):57.
41. Guo Q, Jiang Q, Liu W, Bai Y. All-trans retinoic acid inhibits HOXA7 expression in leukemia cell NB4. Cell Mol Biol. 2016;62(1):636.
42. Liu S, Lei H, Luo F, Li Y, Xie L. The effect of lncRNA HOTAIR on chemoresistance of ovarian cancer through regulation of HOXA7. Biol Chem. 2018;399(5):485–97.
43. McGlinn E, Yekta S, Mansfield JH, Soutschek J, Bartel DP, Tabin CJ. In ovo application of antagomiRs indicates a role for miR-196 in patterning the chick axial skeleton through Hox gene regulation. PNAS. 2009;106(44): 18610–5.
44. Li Z, Huang H, Chen P, He M, Li Y, Arnovitz S, Jiang X, He C, Hyjek E, Zhang J, et al. miR-196b directly targets both HOXA9/MEIS1 oncogenes and FAS tumour suppressor in MLL-rearranged leukaemia. Nat Commun. 2012;3:688.
45. Divoux A, Xie H, Li JL, Karastergiou K, Perera RJ, Chang RJ, Fried SK, Smith SR. MicroRNA-196 regulates HOX gene expression in human gluteal adipose tissue. Obesity. 2017;25(8):1375–83.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.